Passives are not hard to interpret but hard to remember: Evidence from online and offline studies.

Caterina Laura Paolazzi[a]*, Nino Grillo[b], Artemis Alexiadou[c&d], and Andrea Santi[a]

[a] Department of Linguistics, University College London, London, UK; [b] Department of Language and Linguistic Science, University of York, York, UK; [c]Humboldt-Universitat zu Berlin, Institut für Anglistik und Amerikanistik, Berlin, Germany; [d]Leibniz-Center General Linguistics (ZAS), Berlin, Germany.

* Caterina Laura Paolazzi
University College London
Department of Linguistics
Rm 102, Chandler House
2 Wakefield St, London WC1N 1PF

caterina.paolazzi.11@ucl.ac.uk

Passives are not hard to interpret but hard to remember: Evidence from online and offline studies.

Passive sentences are considered more difficult to comprehend than active sentences. Previous online-only studies cast doubt on this generalization. The current paper directly compares online and offline processing of passivization and manipulates verb type: state vs. event. Stative passives are temporarily ambiguous (adjectival vs. verbal), eventive passives are not (always verbal). Across 4 experiments (self-paced reading with comprehension questions), passives were consistently read faster than actives. This contradicts the claim that passives are difficult to parse and/or interpret, as argued by main perspectives of passive processing (heuristic, syntactic, frequentist). The reading time facilitation is compatible with broader expectation/surprisal theories. When comprehension targeted theta-role assignment, passives were more errorful, regardless of verb type. Verbal WM measures correlated with the difference in accuracy, but not online measures. The accuracy effect is argued to reflect a post-interpretive difficulty associated with maintaining/manipulating the passive representation as required by specific tasks.

Keywords: language comprehension; passivization; heuristics; surprisal

Introduction

The representation and processing of passive sentences has puzzled linguists for over 40 years with the big question being: why are passives more difficult to comprehend than actives? (healthy adults – e.g., Ferreira, 2003 –, people with aphasia – Grodzinsky, 1990 –, children acquiring their first language – Maratsos, Fox, Becher & Chalkley, 1985). Broadly speaking, the difficulty is attributed to its non-standard argument order. Standardly, in English, the *doer* (i.e., agent) of an action precedes the *doee* (i.e., patient, see (1)); passives reverse this order (see (2)). The three main proposals for passivization difficulty are: (1) syntactic complexity, (2) heuristics and (3) frequency of use. According to the syntactic complexity account, in passives there is an additional syntactic dependency between where the subject is pronounced and where it is interpreted thematically, increasing complexity. The heuristic account proposes the use of an agent-first heuristic, which then requires revision by syntactic processes. The frequentist approach argues that active sentences are much more frequent in language use, and hence easier to process than passive sentences. While compelling data for any of these accounts is lacking, the contrast in comprehension between active and passive sentences has been used broadly as a measure of "syntactic complexity" across various domains including neuroimaging, language break-down, and language acquisition (Caplan, Waters, Dede, Michaud & Reddy, 2007; Grodzinsky, 1995; Mack, Meltzer-Asscher, Barbieri & Thompson, 2013; Maratsos et al., 1985; Thothathiri, Kim, Trueswell & Thompson-Schill, 2012).

(1) The guitarist$_{\text{AGENT}}$ pushed the singer$_{\text{PATIENT}}$

(2) The singer$_{\text{PATIENT}}$ was pushed by the guitarist$_{\text{AGENT}}$

Ideally the nature of any processing difficulty is understood within the healthy adult population before applying it to other ones. Yet, a review of the passive literature from the healthy adult population presents a scant and heterogeneous picture: passives demonstrate

difficulty on offline tasks that require a judgment of a sentence interpretation (Ferreira, 2003; Street & Dąbrowska, 2010), but no difficulty or even facilitation on online ones that measure the moment-to-moment processing of sentences (Carrithers, 1989; Traxler, Corina, Morford, Hafer & Hoversten, 2014). While the offline data seem consistent with the general tenet that passives are more complex than actives, the online data question it. However, these previous studies collected *either* online or offline measures preventing definite conclusions to be drawn on the possible reason(s) for their contrasting data.

In filling this gap, we present four self-paced reading experiments that simultaneously collected comprehension accuracy data with healthy adults. Results were replicated across 4 experiments, confirming an online vs. offline dissociation and at significance: passives were processed faster than actives at the verb and through much of the by-phrase, but induced more comprehension errors. This picture is inconsistent with the view that passives are more complex than actives. The fourth experiment supports a role for Working Memory (WM) in the accuracy effect. We argue that the complexity observed in offline data are due to post-interpretive processes required of the task and that noncanonical sentences (i.e., passives) are not complex to parse and interpret.

1. 3 Theories of Offline Passive Difficulty

In Ferreira (2003), participants were asked to identify the *doer* (i.e., the agent) or the *acted-on* (i.e., the patient) of an action described in either active or passive sentences. While comprehension accuracy with passives (i.e., 81.5%)[1] was overall high, it was significantly lower than with actives (93.5%). The following sections consider three accounts for this offline difficulty.

Heuristics and the Good Enough Theory:

According to the *Good Enough* (GE) theory (Christianson, Hollingworth, Halliwell and Ferreira, 2001; Ferreira 2003), heuristics are activated in parallel with slower and more precise algorithmic processes, much in the same spirit as Townsend and Bever (2001). Both semantic and syntactic heuristics have been discussed in the literature, but we focus on the syntactic agent-first one, as it is relevant to passivization complexity. According to the agent-first strategy, the first NP of every English sentence is initially interpreted as agent, given this is the most prevalent argument order. The comprehender can only reach the correct interpretation of a noncanonical sentence, if the slower algorithmic processes are given sufficient time and attention to intervene and revise the heuristic. If they are not, then the heuristic can overwhelm the "fragile" syntactic parse, and the heuristic interpretation is deemed "good enough". The data above seems to suggest participants' performance reflects a combination of the two possible outcomes of Good Enough processing. Algorithmic processes most often correct the heuristic, but for some small proportion of trials (~12% of trials) the heuristic is judged "good enough", and left uncorrected by algorithmic processes. The *Good Enough* model has been supported and refined in subsequent studies and review papers by Ferreira, Christianson and colleagues (Christianson, 2016; Ferreira & Patson, 2007; Ferreira & Christianson, 2016, Karimi & Ferreira, 2016).

Syntactic Perspective:

In contrast to Ferreira's model, other mainstream sentence processing models focus on algorithmic processes alone without reference to heuristics. Implications for these processes have arisen from the generative grammar literature, which claim that passives involve a movement operation of the patient/theme argument to the grammatical subject position (e.g., following Chomsky, 1981; Kiparsky, 2013). The additional syntactic dependency in a passive sentence is thought to increase syntactic complexity and tax the parser. An abundance of

evidence demonstrates that other non-canonical, movement-derived structures generate processing difficulty (e.g., object relatives, object clefts; Caplan et al., 2002; Garnham & Oakhill, 1987; Grodner & Gibson, 2005; Santi & Grodzinsky, 2010; Staub, 2010; Traxler, Morris & Seely, 2002). Moreover, this greater syntactic complexity can result in comprehension errors due to processing resources being taxed (e.g., object vs. subject relative clauses; Gibson, 1998, Hakes, Evans & Brannon, 1976). Indeed, the "syntactic complexity" account can explain the comprehension data reported in Ferreira (2003) and has been used to explain the greater brain activation with passive compared to active sentence processing in healthy adults (Mack et al. 2013), delayed acquisition of passives compared to actives in children (Borer & Wexler, 1987) and greater difficulty on passives than actives in aphasic patients (Dickey & Thompson, 2009; Grodzinsky, 1990, 2000).

Frequentist approach

According to he frequentist approach (Johns & Jones, 2015) the more frequently a structure is used by native speakers, the easier its parsing will be. Given that passive sentences are less frequent than active sentences (Gordon & Chafetz, 1990), this model predicts passives to be more complex to process than actives. Studies adopting this approach argue for both online and offline complexity effects for less frequent structures (e.g., object relative vs. subject relative clauses; Reali & Christiansen, 2007; MacDonald, 2013). This theory is also consistent with the data previously reported in Ferreira (2003), but does not fare well under additional data reported therein. No difference in accuracy between subject-cleft and active sentences was observed using the same task. Under a frequency account, the less frequent subject-clefts should demonstrate lower accuracy.

2. Predictions: Online Processing of Passives

The *GE* theory predicts that passive sentences should be overall slower to read than actives. Algorithmic processes often intervene to correct/override the initial heuristic interpretation, which should be reflected in a slow-down relative to active sentences where no correction/adjudication is required. This revision/adjudication can only occur once the algorithmic processes have identified and resolved the syntactic dependency between the verb and its internal argument (i.e., filler-gap dependency). The earliest point where this can occur is at the verb, as other canonical structures are still possible prior to that point (e.g., "The man was visiting the woman"). A previous Cross Modal Lexical Priming (CMLP) study (e.g., Osterhout & Swinney, 1993) suggests that "filler-gap" dependency resolution in passives is delayed relative to the gap by 500ms (and up to 1000ms for a reliable effect). Thus, the revision effect may also be downstream from the verb (i.e., within a couple of words of the verb). The syntactic account likewise predicts the difficulty effect to arise at the point of dependency resolution: at the verb (or shortly thereafter). The frequentist account does not have a locational prediction, but does predict passives to be read slower than actives.

Current online evidence, however, does not support these predictions (Carrithers, 1989; Traxler et al., 2014). In contradiction to *GE, Syntactic Complexity* and *Usage-based* theories, passive sentences are read faster than active sentences. In a subject-paced, word-by-word reading task, Carrithers (1989) found that passive sentences were read over 20 ms per word faster than active sentences, and the difference appeared "after the first noun phrase had been processed" (p. 80), although neither a precise location nor example sentence is provided, making this result difficult to interpret. Traxler et al. (2014) also used self-paced reading and found that passive sentences (see (3) below) were read numerically, but not significantly, faster than active sentences (see (4) below) at the verb (i.e., "tricked") and object NP (i.e., "cowboy").

(3) The farmer was tricked by the cowboy into selling the horse.

(4) The farmer tricked the cowboy into selling the horse.

While the data from both studies appear inconsistent with GE (Ferreira, 2003; Karimi & Ferreira, 2016; Ferreira & Christianson, 2016; Ferreira & Patson, 2007), Syntactic Complexity and Frequentist approaches, no comprehension accuracy data were collected/reported by them to provide a complete picture. It is possible these participants relied more on the heuristic and hence no revision was observed. It could also be that the by-phrase may not have been sufficiently long to detect a complexity effect in these studies. As mentioned, dependency resolution is reported downstream from the verb in passives when using the CMLP paradigm.

Potentially consistent with the faster reading times of passives than actives are expectation-based (e.g., Levy, 2007) or surprisal-based accounts (e.g., Hale, 2001). According to these models, sentence processing unfolds in a parallel, incremental and probabilistic fashion. In other words, the relative difficulty of processing upcoming material is dependent on the expectations created by the current representation of the sentence. Previous data in support of these theories come from various sources (Konieczny, 2000; Staub, 2010).

The morphological richness of the English passive sentence contributes to expectations for upcoming syntactic categories. Unlike actives, passives have an additional auxiliary and a(n) (optional) "by". The presence of an auxiliary following a subject NP increases the expectation for a verb compared to a subject NP alone. Likewise, the "by" following a past participle increases expectations for a determiner (DP) with respect to a past participle alone. These increased expectations decrease processing demands and speed up reading time.

Despite surprisal and expectation-based theories being consistent with the reading time data for passivization, they do not predict the offline results reported by Ferreira's (2003) listening study. It would thus be necessary to find an alternative explanation to account for the offline complexity effect. The next section considers a potential explanation: verbal Working Memory (vWM).

3. The role of verbal working memory in passives processing.

Verbal Working Memory (vWM) provides a temporary store for a relatively small amount of phonological information while further linguistic input is processed (see Caplan & Waters, 2013, for an extensive review of studies on memory mechanisms in language comprehension). It should not be surprising that vWM has been shown to correlate with subjects' performance on a wide variety of offline linguistic tasks that indirectly tap into the degree of successful sentence interpretation (e.g., acceptability judgments, verification task; Boyle, Lindell & Kidd 2013; Conway et al., 2005; Daneman & Carpenter, 1980; DeDe, Caplan, Kemtes & Waters, 2004; Jaeggi, Buschkuehl, Perrig & Meier, 2010; Kim & Christianson, 2013; Roberts & Gibson, 2002; Sprouse, Wagers, & Phillips, 2012; Swets, Desmet, Hambrick & Ferreira, 2007). Interestingly, it has not been found to correlate with online language processing, or only in limited circumstances (Caplan, DeDe, Waters, Michaud & Tripodis, 2011; Evans et al., 2014).

The above discrepancy between offline and online data and its correlation with vWM bears some semblance to what has been discussed for passivization difficulty: it is observed offline, but not online. This suggests vWM scores may also correlate with offline comprehension accuracy with passive sentences. We consider two reasons for this. First, the *Good Enough* model of passive sentence processing predicts individuals with a lower vWM span to rely more on heuristics, as algorithmic processes pose greater demands on the parser (Christianson, Williams, Zacks & Ferreira, 2006; Karimi & Ferreira, 2016). Second, individuals with a lower vWM span might have more difficulty maintaining the passive representation in memory to answer a following comprehension question. The comprehension task may be more demanding on memory mechanisms in the case of noncanonical structures, if further manipulations in memory are required. In either case, a positive correlation between vWM span and accuracy to comprehension questions is expected (i.e., the lower the vWM span the less accurate they should be).

If indeed such a correlation were to be observed, then we could hypothesize that parsing and interpreting a passive sentence is not difficult *per se* (as suggested by the faster reading times in online tasks). Rather, passives rely more on heuristics and/or operating on or maintaining its full representation is more difficult than for an active.

4. Predicate x Passivization Interaction: Overlooked by Previous Studies

Our review of previous psycholinguistic studies also indicated that the interaction between passivization and predicate type for interpretation and availability was overlooked. While eventives consistently deliver verbal passives (and hence require movement) across languages, passives of states can be temporarily interpreted adjectivally at least until the by-phrase is introduced (e.g., "John is (very) cherished" has an adjectival interpretation, while "John is cherished by Mary" a verbal one), in English. Further, in English, not all states can passivize, but those that can, are also those that can be coerced into a state consequent to an event. Hence, passives of states have been proposed to require coercion of a state consequent to an event (Gehrke & Grillo, 2009). Passives with predicates producing a result or change of state (e.g., "to kick" or "to push") seem to be acquired earlier than passives of other types of predicates (e.g., stative verbs, like "to love"; Maratsos et al., 1985; Volpato, Verin & Cardinaletti, 2013; but see Messenger, Branigan, McLean & Sorace, 2012, for an interesting discussion about the discrepancy in results observed across different tasks). Paralleling the data reported in the acquisition literature, aphasic patients' performance on passivization varies with the predicate type tested: passive sentences containing stative predicates were in fact found to be more difficult to produce (e.g., lower number of sentences produced in elicited production task) and understand (e.g., more errors in sentence verification task) for aphasic patients than sentences containing eventive predicates (Grodzinsky, 1995). During completion of the present manuscript, evidence in support of stative predicates increasing passivization difficulty was

also reported in healthy adults' acceptability judgments (Ambridge, Bidgood, Pine, Rowland & Freudenthal, 2016).

Passivization of states may be more errorful/costly than passivization of events, due to the difficulty of representing the eventive reading of a stative predicate that requires resolving the temporary ambiguity and coercion. Given that in previous studies predicate type was not controlled for (Carrithers, 1989; Ferreira, 2003; Traxler et al., 2014), the results are further difficult to interpret, due to the possibility that they represent a mix of effects.

5. Current studies: Aims and Predictions

The current study aimed at filling a gap in the literature on passive sentence processing in healthy adults by simultaneously collecting comprehension accuracy and reading time (self-paced) data while manipulating the predicate's event structure.

Experiment 1 only contained eventive predicates, to maximize the possibility of detecting a complexity effect of syntactic movement. However, we found no online complexity effects, rather passives were read significantly faster at the verb and regions early in the by-phrase. We also found no offline comprehension effects. Experiment 2 only contained stative predicates to investigate possible effects of temporary ambiguity and/or coercion. This study replicated the significantly faster online reading data for passives, but found greater offline errors for passive sentences. Experiment 3 directly investigated a possible interaction between syntax and predicate type in a within participant design. We failed to find an interaction between syntax and predicate type, but simply observed more comprehension errors on passives along with faster reading times. Given that earlier studies' found vWM correlates with offline accuracy but not online reading times, Experiment 4 investigated whether the offline results could be explained by vWM demands. In (partial) support we found a correlation between vWM capacity and the accuracy difference between active and passive sentences.

Collectively, the data argue against *Syntactic Complexity*, *Usage-based* and *Agent-first Heuristic* theories. The faster reading times online are consistent with expectation-based and surprisal-based accounts, where greater morphological cues in the passive than active results in increased expectations for upcoming words (or categories at the verb and early in the by-phrase). These data are also in line with results collected using other methodologies (visual world paradigm; Kamide, Scheepers & Altmann, 2003), which show that there is no cost associated with interpreting passives online. Although an initial preference for an active interpretation might be at play (it is unclear whether the use of the agent-first heuristics is driven by the Visual World paradigm), this is immediately corrected within the verb region itself, without a processing cost. We use these data to argue that the offline difficulty effect likely reflects task-related post-interpretive processing, but that passives are not inherently more complex to parse and interpret.

Experiment 1

This first study sought to find comparable offline and online difficulty effects for the passive vs. active contrast in line with the two main theoretical accounts in the literature. In order to maximize the possibility of detecting an effect of syntactic complexity we used only eventive predicates, as they are interpreted verbally in the passive, and hence provide the cleanest test for movement complexity.

Methods and Design

1. Participants

Thirty-five native British English-speakers were recruited to participate in the study (24 females; average age: 28.6). They were all aged between 18 and 50 and had no visual or hearing impairment.

Participants were recruited through the UCL Sona System and received either payment or course credits for their participation. All the participants were informed of the aims and procedures of the experiment and provided informed consent, approved by UCL ethics.

2. *Materials*

There were three conditions (see Table 1 for example sentences): (1) active perfect, (2) active progressive, and (3) passive. The progressive was included to act as an additional control, which is matched to the passive for the auxiliary (e.g., "was"). Thirty sentence sets were generated. The sentences all contained eventive predicates. The by-phrase/direct object noun was modified by two conjoined adjectives (pre-nominally) in order to allow sufficient time to detect an online complexity effect in passives, as previous cross-modal lexical priming studies show delayed reactivation of the filler in passives (Osterhout & Swinney, 1993) unlike wh-dependencies which show immediate reactivation at the gap (Love & Swinney, 1996). The sentence final preposition phrases were included to avoid any end-of-sentence effects.

The length (i.e., number of words per sentence) was kept constant within each condition: given the syntactic differences across conditions, passive sentences always had 2 words more than simple actives (i.e., auxiliary and "by"), and progressive sentences always had one word more than simple actives (i.e., auxiliary; see Table 1 for example sentences; for the complete list of items, see Appendix 1). Additionally, 60 filler sentences, with varying complexity (15 actives; 15 passives; 15 sentences with negation; 15 garden-path "while…" constructions), were created to mask the purposes of the experiment (for the list of fillers, see Appendix 2). [Table 1 near here]

Pre-norming: Plausibility. The order of the NPs was not reversed across active and passive items. This means that the NPs in passive sentences were assigned the reversed thematic roles with respect to active sentences. To ensure that plausibility was not affected by reversing the

theta-role NP combinations across conditions, the experimental items were normed in 2 tasks administered via online questionnaires designed in Qualtrics (https://www.qualtrics.com/). 72 participants (recruited online through the UCL Subjects Pool https://uclpsychology.sona-systems.com/Default.aspx?ReturnUrl=/; 41 females; mean age: 26.84) rated, on a 7-points scale from highly implausible (e.g., 1) to highly plausible (e.g., 7), the plausibility of the experimental items and of implausible items created ad hoc.

The first questionnaire targeted the plausibility of the two thematic role assignments of each item. Both orders were tested in the active form without the PPs. Experimental items were thus NP-VP-NP active sentences, and the manipulation consisted in reversing the NPs. The second questionnaire targeted the plausibility of the entire sentence across the progressive and passive condition to further ensure the selected predicates could be conjugated in the progressive form without affecting the overall plausibility of the sentence. Implausible items had the same structure of the experimental sentences, and hence changed accordingly from the first to the second questionnaire. The implausible items resulted from argument role reversals (see (5)), or general semantic/pragmatic oddness (see (6); for the complete list of implausible items, see Appendix 3).

(5) The law abiding police man was arrested by the criminal last Saturday on the town's busy high street.

(6) Santa Claus gave coal to all the children on the nice list and presents to those on the naughty list.

Data collected in the first questionnaire were analysed using a linear mixed effects model, containing the order of theta-roles as fixed effect and both subjects and items as random effects (including both intercepts and slopes). The contrasts used were passive order (e.g., "The attractive and talented singer rejected the guitarist.") vs. active order (e.g., "The guitarist

rejected the attractive and talented singer."; [0.5, -0.5]). P-values were determined through treating the t-value as a z-statistic (Barr, Levy, Scheepers, & Tily, 2013). The analysis revealed that there was no significant difference between the two possible orders of arguments (average active: 4.84; average passive: 5.04; β=-.2, t=-.16, p= .25).

Data collected in the second questionnaire were analysed similarly to data collected in the first questionnaire, but the fixed effect was syntax and the contrast used was passive vs. progressive [0.5, -0.5]. No significant difference was found between the two conditions (average progressive: 4.83; average passive: 4.94; β=-.19, t=-.88, p= .38). Hence, the items did not significantly differ either in the plausibility of thematic role assignment across conditions, or in the overall plausibility of the sentences across the progressive and passive condition.

BNC Corpus Analysis: Structural Frequency of the Verbs. Finally, an analysis of all the verbs' entries contained within the British National Corpus was conducted (BNC; http://www.natcorp.ox.ac.uk/) to exclude possible frequency interpretations of our results. Using the "Phrases in English" tool provided by the BNC, the first 100 instances of each verb in its verbal past-tense form (i.e., as used in the experiment, not as an adjective) were selected. The verbs were analysed in their original sentential context in order to categorize each instance as being in the active or passive voice. An analysis on the frequency of their surface form found the verbs to be more frequent in the active (see Appendix 4 for a table containing the frequency of surface forms of all our verbs).

3. *Procedure*

The normed items (90 experimental sentences) were presented in a non-cumulative self-paced reading paradigm using Linger 2.88 software (http://tedlab.mit.edu/~dr/Linger/readme.html).

Verification questions requiring a "yes" or "no" button response followed each item, (experimental and filler) to ensure participants' active comprehension during the task (for the

list of comprehension questions, see Appendix 5). Questions were designed to be half correct and half incorrect. Questions could either be simple (e.g., focusing on various attributes within the sentence, with an example of such a question to the sentence in (1) being "Did the musician play the piano?") or complex (e.g., targeting theta-role assignment). The rationale was to try to target all aspects of the sentence to avoid strategic processing.

Sentences were presented with the words masked by dashes. An empty space appeared between words. A press of the space bar unmasked one word at a time. On the contrary, comprehension questions were presented unmasked on the screen and participants pressed either "j" for "no", or "f" for "yes". A gaming keyboard (Razer® Blackwindow) was used for accurate button press timing. Participants were provided with feedback if they chose the wrong answer. Practice trials (6 in total) allowed them to familiarize with the task prior to the testing session. The task lasted approximately 30 minutes and was administered in a soundproof room.

4. Data analysis

Due to low accuracy on fillers (lower than 75% overall), 5 participants were excluded from analysis. Hence, data from 30 participants were analysed.

Reading Time and Question Response Time Data. The outcome measures were: accuracy and reaction times to verification questions and reading times. The analysis was run using RStudio, an application for data analysis (https://www.rstudio.com/). Unreasonably high (>2500ms) and low (<100ms) raw reading times were excluded. Residual logRTs were calculated based on word length and the restricted cubic spline of word position (Hofmeister, 2011) considering all sentences (experimental and filler). The residuals were then analysed to identify further possible outliers: data above or below 2.5 times the standard deviation from the mean (by subject, condition, and region) were excluded. This resulted in 0.24% of the original data being

removed. In terms of response times, unreasonably high (>12000ms) and low (<700ms) values were removed. This resulted in 0.14% of the original response time data being removed.

The cleaned residual logRT reading and logRT reaction time data were analysed using a linear mixed effects model including syntax as a fixed effect and both subjects and items as random effects (including both intercepts and slopes). The contrasts used were passive vs. actives (both progressive and simple active) [2/3, -1/3, -1/3], and progressive vs. simple active [0.5, -0.5]. P-values were determined through treating the t-value as a z-statistic (Barr et al., 2013). This model was run for each of the following regions of interest for the reading time data: (1a) auxiliary, (1b) verb, (2) determiner of the by-phrase, (3) first adjective, (4) conjunction, (5) second adjective, (6) object NP and (7&8) 2 words after the object NP. Only the analysis of reading times in accurate trials will be reported and discussed in the Results section, given that this is critical for the predictions of the *Good Enough* theory (revision would need to take place only on accurate trials). However, both accurate only and all (accurate and inaccurate) trials were analysed separately and compared, and no significant difference was found between the two analyses.

Comprehension Accuracy Data. Accuracy data were analysed using a mixed effects logistic regression with a binomial distribution including the same effects and contrasts as the analysis of the online data. The full model (containing both intercepts and slopes for random effects) was always run initially, but when convergence could not be met using the full model, the models were modified to meet convergence (the Results section describes when these modifications were needed and what they were).

Results

Comprehension Question Results:

Offline accuracy and RTs did not reveal significant differences across conditions. Accuracy to comprehension questions was almost identical across conditions (passive: 85.71%; progressive: 83.67%; simple active: 83.3%; passive vs. actives: β=.04, z=.2, p=.84; simple active vs. progressive: β=.03, z=-.11, p=.9; see Table 2[2]), and, similarly, RTs did not significantly differ across conditions (passive vs. actives: β=.004, t=.27, p=.79; simple active vs. progressive: β=.008, t=.52, p=.6; see Table 2[3]). Given that Ferreira (2003) argues that comprehension of passive sentences is selectively impaired with questions targeting theta-role assignment, we separately analysed the participants' performance on theta-roles questions only. The analysis did not reveal any significant difference with respect to the overall analysis: participants' performance was generally lower on theta role questions than on other question types (passive: 78.5%; progressive: 81.12%; simple active: 79.4%), but accuracy did not significantly differ across conditions (passive vs. actives: β=-.11, z=-.45, p=.66; simple active vs. progressive: β=-.05, z=-.15, p=.88). Similarly, RTs to theta-role comprehension questions did not differ from RTs to other question types and there was no significant difference across conditions (passive vs. actives: β=.01, t=.57, p=.57; simple active vs. progressive: β=.02, t=.77, p=.44). In order to exclude possible learning effects that might have confounded our results, we ran and compared 2 separate analyses on the data collected in the first vs. second half of the sessions[4]. There was no significant difference between the results of the two (first vs. second) analyses. It should be noted that the active sentences have a lower accuracy with respect to results reported by previous studies (Ferreira, 2003). This is not surprising given our sentences were significantly longer (long PPs and prenominal modification) and our comprehension questions assessed all aspects of the sentence. In combination, they provided a more demanding task for the participant that is bound to reduce accuracy. Overall, the offline data provide no evidence for passive sentences being more difficult to understand. [Table 2 about here]

Reading Time Results:

Analysis of reading times on correct trials only revealed that the verb was read significantly faster in passive sentences, with respect to active sentences (β=-.03, t=-2.42, p=.01) and in progressive sentences with respect to simple active sentences (β=.05, t=2.54, p=.01). This outcome is likely related to the presence of the auxiliary in both the passive and the progressive conditions, which eases processing at the following verb due to a smaller *surprisal* effect, i.e., a smaller cognitive load in processing the subsequent word, given its high probability to follow in the sentence (Hale, 2001). In fact, the auxiliary both creates a strong expectation for a verb to follow and determines tense. In the active, tense must be computed at the verb, thus additionally slowing processing with respect to the other two conditions.

Up to 4 regions after the verb, passive sentences were read numerically faster than actives and significantly so at 3 of them (at the determiner: β=-.04, t=-2.75, p=.006; at the conjunction: β=-.04, t=-3.43, p=<.001; at the second adjective: β=-.03, t=-2.07, p=.04). No significant difference was found at the first adjective or after the second adjective. Results are presented in Figure 1. Finally, the same first vs. second half analysis described for our offline data was conducted on the online data. Again, no significant difference was observed with respect to the overall results. [Figure 1 about here]

Discussion

The results from the first study, which simultaneously collected online and offline measures of passivization difficulty, can be summarised in three main points: (1) passive sentences were consistently read faster than actives at the verb and much of the by-phrase, indicating that they are not more difficult to process; (2) passive sentences were comprehended as accurately as active sentences, indicating that they were not harder to interpret.

Overall, these findings are inconsistent with a heuristic processing of passive sentences, at least in Ferreira's (2003) terms of a commitment to the agent-first strategy that would later require revision of the initial incorrect interpretation. As noted, this would predict passives to be processed more slowly than actives when the algorithmic processes revise/adjudicate with the incorrect heuristic interpretation.

Likewise, if passivization is obtained via movement (Chomsky, 1981; Kiparsky, 2013), we do not find a processing cost associated with it.

Finally, these data are also incompatible with a usage-based approach to language processing which would predict longer reading times for passives than active, given they are used less frequently (Johns & Jones, 2015).

Rather the results are in line with other studies, which report that processing noncanonical sentences is not necessarily difficult, but rather dependent on other factors, such as the nature of the material intervening in the movement dependency. For example, longer reading times in object- vs. subject-extracted relative clauses/clefts were found to be dependent on the syntactic/semantic similarity between the NPs in the sentences (Gordon, Hendrick & Johnson, 2001; Gordon, Hendrick, Johnson, & Lee, 2006; Gordon, Hendrick, & Levine, 2002). When the similarity was reduced, no (or less) reading time difference was observed.

The faster reading times observed in passives than actives could be explained by surprisal (Hale, 2001). As already mentioned in the introduction, in English, the passive structure is morphologically richer (e.g., auxiliary, past participle morphology) than the active structure, which could provide greater expectations for upcoming syntactic constituents.

The offline data are also compatible with surprisal accounts. Moreover, the consistency between offline and online data in our results suggests an identical underlying explanation. However, the question that remains is why do passives generate worse accuracy in studies like Ferreira (2003)? Since the current experiment focused on eventive predicates, it may be that

stative predicates are the main contributor to passive difficulty, given their temporal ambiguity and/or need for coercion. The use of mixed predicate types in previous studies (e.g., Ferreira, 2003) might have induced more comprehension errors in passive sentences due to the stative predicates. Experiment 2 tests this hypothesis by using the experimental design with only stative predicates.

Experiment 2

The second experiment aimed at testing whether the greater complexity of passives, previously observed in accuracy, emerged from the stative predicates. The design resulted from a single but fundamental modification of the first experiment design: the predicates were always stative, and particularly subject-experiencer psych predicates. Under temporal ambiguity, the parser may favour the simple adjectival interpretation, which then requires revision. This ambiguity or greater complexity of two parses may lend itself to more offline errors. Thus, we expected passive sentences in Experiment 2 to be understood less accurately than actives. They may also be read more slowly if there is need for revision after the ambiguity is resolved.

Methods and Design

1. Participants

Twenty-six native British English-speakers were recruited to participate in the study (21 females; average age: 23.4).

The same recruitment criteria and procedures were used as in Experiment 1. None of the participants had previously participated in any of the experiments related to this project (including the pre-norming tests).

2. Stimuli

Three modifications, related to the experimental manipulations, had to take place: (1) complex event predicates were substituted with subject experiencer predicates; (2) the locative PPs that are not acceptable with subject experiencer predicates were substituted with implicit causal clauses; (3) the progressive condition was omitted as the progressive is not compatible with subject experiencers. To avoid an auxiliary bias, (i.e., only encountering a passive following the auxiliary "was"), 10 of the filler sentences of Exp. 1 were modified to include "was" with the verb in the progressive form. Finally, some of the argument pairs were slightly modified to adjust to the new verbs and create plausible experimental items. Everything else (types of filler sentences; overall sentence length) was kept identical to Exp. 1. 30 sentence sets were generated as in Exp. 1. Examples of experimental items are presented in Table 3 (for the complete list of items, see Appendix 6). [Table 3 near here]

Pre-Norming: Plausibility. The experimental items were then normed in a plausibility study administered via an online questionnaire designed in Qualtrics (https://www.qualtrics.com/). Participants for this study were recruited via Prolific Academic, a platform for online research (https://prolific.ac/). Filler implausible items, rating scale, structure of experimental items (only active sentences; manipulation: order of arguments) were identical to the first plausibility questionnaire run in Exp. 1. 66 participants (37 females; average age: 31.5) were recruited to participate in the online questionnaire. The results of the plausibility task revealed that there was no significant difference between the two possible orders of arguments (average active: 5.89; average passive: 5.76; β=-.02, t=.71, p=.48). Hence, the items did not significantly differ in the plausibility of thematic role order.

BNC Analysis: Frequency of Verb in Active-Passive Surface Form. Finally, an analysis of all the verbs' entries contained within the British National Corpus (BNC; http://www.natcorp.ox.ac.uk/) was conducted following the same procedure as in Exp. 1.

3. *Procedure*

The same procedure as in Exp. 1 was used (see Appendix 7 for the complete list of comprehension questions used in Exp. 2).

4. *Data analysis*

Due to low accuracy results on fillers (lower than 75% overall), 2 participants were excluded from the final analysis. Hence, data from 24 participants were analysed.

The outcome measures were: accuracy and reaction times to comprehension questions and reading times. The analysis was run using RStudio, an application for data analysis (https://www.rstudio.com/).

The analysis of reading times, accuracy and Reaction Times data followed the same steps as in Exp.1. 0.31% of the original response time data were removed.

Results

Comprehension Question Results:

In contrast to Exp. 1, comprehension question measures differed across conditions. Questions following passives were found to be responded to significantly slower (passive vs. active: β=.04, t=2.49, p=.01; see Table 4) and significantly less accurately (passive: 78.3%; active: 86.1%; passive vs. active β=-.61, z=-2.941, p=.003; see Table 4) than those following actives. Given Ferreira's (2003) hypothesis that passivization difficulty is selective to theta-role questions, we separately analysed the participants' performance on these questions. The analysis revealed the same significant effect: participants' performance on theta role questions was generally lower than for other question types (passive: 72.2%; active: 82.87%), and accuracy significantly differed across conditions (passive vs. actives: β=-.58, z=-2.4, p=.02). Similarly, there was a significant difference across conditions in Reaction Times to theta-role

comprehension questions only (passive vs. actives: β=.06, t=2.9, p=.004). In order to exclude possible learning effects that might underlie our results, we ran and compared 2 separate analyses on the data collected in the first vs. second half of the sessions[5]. The analysis did not reveal any significant difference with respect to the overall results. Finally, we believe that the low accuracy to active sentences, with respect to results reported by previous studies (Ferreira, 2003), was determined by our longer stimuli with respect to previous designs (see Results section of Experiment 1 for further detail).

These data indicate that passives were indeed harder to understand than actives with stative predicates. [Table 4 about here]

Reading Time Results:

Analysis of reading times[6] on correct trials only revealed that passive sentences were read numerically faster than active sentences up to the 2nd adjective, but only significantly at the determiner (β=-.04, t=-3.18, p =.001). The reverse effect was observed after the second adjective (a marginally significant difference was observed at the head of the by-phrase: β=.04, t=1.72, p=.08). No numerical trend could be observed after the 4th region following the verb. Results are presented in Figure 2. Finally, the same first vs. second half analysis described for our offline data was conducted on the online data. Again, no significant difference was observed with respect to the overall results. [Figure 2 about here]

Discussion

The results of our investigation of passivization with stative predicates can be summarised in the following two points: (1) passive sentences were comprehended less accurately than active sentences, and (2) passive sentences were read faster at the determiner of the by-phrase. There was a reversal of this reading time difference at the head noun, but the

result was only marginally significant. The verb was again read numerically faster in the passive than the active, but this effect was not significant in the current experiment.

Considered together with the data obtained in Exp. 1, the results form a very interesting picture: passive sentences are processed faster than active sentences online, regardless of the predicate type, consistent with the expectation-based account. However, offline, passivization does seem to interact with the predicate type: more specifically, passivizing a stative predicate creates more difficulty than passivizing eventive predicates. The present results then raise at least two questions: (1) why are passives processed faster than actives online? and (2) what causes the difficulty in interpreting a passivized stative predicate?

A first attempt to interpret these data may be to apply a speed-accuracy trade-off analysis. However, the results are not compatible with this explanation for three reasons: (1) we analysed reading times from accurate trials only and still found passives to be significantly faster than actives; (2) participants took longer to answer comprehension questions regarding a passive rather than an active sentence and still made more errors on the former than the latter; (3) in Experiment 1 we also saw faster reading times for passives but without any accuracy difference.

At present, the only processing models compatible with our online results are expectation-based or surprisal-based accounts of syntactic comprehension (e.g., Hale, 2001; Levy, 2007). Just as in the first experiment, neither plausibility (excluded by a pre-norming study) nor frequency of surface form (excluded by a post-experiment analysis of the BNC; the table containing the frequency of surface forms of all our verbs is contained in Appendix 8) can explain the faster reading times.

On the other hand, the accuracy findings suggest an interaction between the predicate properties and passivization when it comes to interpretation difficulty. The difficulty interpreting passives of states could result from two factors: (1) the temporary ambiguity

generated in English between adjectival and verbal passive; (2) the required coercion of the verb meaning to allow for the verbal interpretation. The first factor would require a revision of the initial incorrect adjectival interpretation and its effect could be signalled by the longer reading times in passives with respect to actives at the head of the by-phrase that was marginally significant.

However, the presence vs. absence of a significant difference in accuracy across Experiment 1 and 2 does not represent evidence in favour of an interaction between passivization and predicate semantics. Hence, in order to strengthen our claim, i.e., that passivization and predicate structure interact in processing, we will run a third experiment containing the same stimuli as Exp. 1 and 2, but with a within-subject, 2x2 design that manipulates syntax and predicate type. Moreover, we will use a more sensitive test for interpretation difficulty and focus all comprehension questions (for experimental items) on thematic role assignment, as this is the most crucial question for assessing the agent-first heuristic of the Good Enough theory.

Experiment 3

The third experiment was mostly a replication of Exp. 1 and 2, but using a within-subject design. Both the syntax of the sentence (passive vs. active) and the predicate type (eventive vs. stative) were manipulated. To provide a more sensitive test of accurate thematic role assignment we only used the comprehension questions that targeted thematic role assignment (with experimental items). Both theta-role and simple questions continued to appear after filler items with an increase in proportion of the simple questions to keep proportions comparable to Exp. 1 and 2. The motivation for focusing experimental item questions on theta roles was to have a more comparable design to previous experiments that found passives errorful in terms of interpretation (e.g., Ferreira, 2003) and increase any chance of observing a difficulty effect

of passivization on interpretation. If indeed the difficulty in processing passives arises from the incorrect application of a heuristic strategy that interprets the first NP as agent, as argued by Ferreira (2003), then the interpretation of thematic roles should be selectively disrupted.

Given the results of the previous experiments, the predictions of Exp. 3 were that: (1) passives should be processed significantly faster than actives online at different points within the by-phrase (i.e., we should observe faster reading times in passives with respect to actives), but possibly slower at the head of the by-phrase in statives only, signalling revision of the initial incorrect adjectival interpretation, as was suggested from Experiment 2; (2) based on the results from Exp. 1 and Exp. 2 we expected to observe an interaction between the syntax of the sentence and the predicate type offline (i.e., in accuracy and reaction times to comprehension questions). Specifically, we expected passives to be responded to less accurately than actives for stative predicates, but no (or smaller) difference with the eventive ones. Alternatively, if our previous test of comprehension was not sufficiently sensitive and/or the difference in differences was not significant, then we would simply see a main effect of passivization (i.e., more errors on passives regardless of predicate type).

Methods and Design

1. Participants

Sixty-five native British English-speakers were recruited to participate in the study (46 females; average age: 23).

The same recruitment criteria and procedures were used as in Experiment 1-2. None of the participants had previously participated in any of the experiments related to this project (including the pre-norming tests).

2. Stimuli

Twenty-eight sentence sets were chosen among the sets used in Exp. 1 and 2. Each sentence set contained 4 sentences, 1 per condition. The sentences followed a 2 syntax (active vs. passive) by 2 predicate type (eventive vs. stative) design. To avoid the confound of an auxiliary bias, (i.e., only encounter a passive following the auxiliary "was"), 10 of the filler sentences of Exp. 1 were modified to include an auxiliary and a verb in the progressive form. Everything else (types of filler sentences; overall sentence length) was kept identical to Exp. 1 and 2. Examples of experimental items are presented in Table 5 (for the complete list of items, see Appendix 9). [Table 5 near here]

Pre-norming: Plausibility. The pre-norming plausibility test was administered via an online questionnaire designed in Qualtrics (https://www.qualtrics.com/). Participants for this study were recruited via Prolific Academic, a platform for online research (https://prolific.ac/). Filler implausible items, rating scale, structure of experimental items (only active sentences; manipulation: order of arguments) were identical to the plausibility questionnaires run in Exp. 1 and 2. 44 participants (25 females; average age: 30.5) were recruited to participate in the online questionnaire. Data were analysed using a linear mixed effects model, containing the order of theta-roles as fixed effect and both subjects and items as random effects (including both intercepts and slopes). The contrasts used were passive order vs. active order [0.5, -0.5]. P-values were determined through treating the t-value as a z-statistic (Barr et al., 2013). Based on the plausibility ratings, 28 sets that displayed no significant difference between the two possible orders of arguments (average active: 5.41; average passive: 5.40; β=.01, t=.09, p=.92) were chosen. Hence, the items did not significantly differ in the plausibility of thematic roles order. However, the items in the stative conditions were rated significantly more plausible (β=.31, t=2.46, p=.01) than the ones in the eventive conditions. Plausibility did not differ according to syntax or to an interaction between syntax and predicate type. Despite not undermining the possibility of analysing active vs. passive across conditions or their interaction

with verb type, this implies that we could not directly compare results across predicate types alone.

3. *Procedure*

The same procedure as in Exp. 1 and 2 was followed. However, the comprehension questions following each experimental item only targeted theta-role assignment (e.g., "Did the musician reject the guitarist?"; for the list of comprehension questions, see Appendix 10). To avoid creating a bias in attention towards specific parts of the sentence (in this case, the NP-VP-NP part), fillers were followed by complex and simple questions, as in previous experiments.

4. *Data analysis*

Due to low accuracy results on fillers (lower than 75% overall), 5 participants were excluded from the final analysis. Hence, data from 60 participants were analysed.

The analysis of reading times, accuracy and Reaction Times data followed the same steps as in Exp.1. and 2. 0% of the original response time data were removed.

Results

Comprehension Question Results:

The offline results did not demonstrate an interaction as suggested by the results of Exp. 1 and 2. In comprehension accuracy we found an effect of syntax (β=-.44, z=-2.08, p=.04) due to accuracy being lower following a passive rather than active sentence and an effect of predicate type (β=-.29, z=-2.09, p=.04), due to lower accuracy following a stative rather than eventive predicate (see Table 6[7]). The interaction between predicate and syntax in accuracy was not significant, as was expected from Experiment 1 and 2. The Reaction Times to comprehension questions showed a significant effect of syntax (β=.06, t=4.26, p<.001), due to RTs being longer following a passive sentence than its active counterpart (see Table 6). The predicate type

effect on reaction times was not significant nor was the interaction. Overall, the data indicate that passive sentences were more difficult to interpret than active sentences, and difficulty was greater when the predicate was a stative rather than an eventive verb. However, we did not find a direct interaction between syntax and predicate type, contrary to what we expected. [Table 6 about here]

Given this discrepancy, it is worth noting that unlike Experiment 1 and 2, we observed a large amount of variance in the offline data. The average variance was largely affected by participants performing at chance on actives (15 out of the 60 analysed) and on passives (21 out of the 60 analysed). Again, this performance was unlike what we observed in Exp. 1 and 2, where no participant performed at chance on actives and only one participant in Exp. 1 (out of the 30 analysed) and one in Exp. 2 (out of the 24 analysed) performed at chance on passives. We will return to contemplate the underlying source of these differences in the Discussion section.

Finally, in order to exclude possible learning effects that might have confounded our results, we ran and compared 2 separate analyses on the data collected in the first vs. second half of the sessions[8]. In accuracy, differently from the overall analysis, the syntax effect was only observed in the second half, and not first half, of the sessions, which contradicts a learning effect interpretation of this discrepancy. Moreover, in Reaction Times to comprehension questions, the syntax effect was significant across both sessions, indicating, once again, that the absence of a syntax effect in accuracy in the first half of the sessions cannot be due to learning effects.

Reading Time Results:

The analyses of reading times on correct trials only replicated the main results of Exp. 1 and 2. In fact, we found a significant effect of syntax at the verb (β=-.06, t=-2.68, p=.007[9]), determiner

(β=-.05, t=-6.07, p<.001), first adjective (β=-.02, t=-2.08, p=.04), and conjunction (β=-.02, t=-2.56, p=.01; see Figure 3), indicating that passives were read faster than actives in these regions. No complexity effect was found at the head of the by-/object phrase in passive with respect to active sentences, contrary to what was reported in Exp. 2. No effect of predicate type or interaction between syntax and predicate type were found. Overall, the data further confirm that passive sentences are processed faster than active sentences. Finally, the same first vs. second half analysis described for our offline data was conducted on the online data. No significant difference was observed with respect to the overall results. [Figure 3 about here]

Discussion

Experiment 3 found passives to be processed faster than actives, but accuracy to comprehension questions to be lower for passives than for actives. There was no interaction between passivization and predicate type in accuracy (or reading times). As per Experiment 2, this contrast between online and offline data is not compatible with a speed-accuracy trade-off account for two reasons: (1) we analysed reading times from accurate trials only and still found passives to be significantly faster than actives; (2) participants took longer to answer comprehension questions regarding a passive rather than an active sentence and still made more errors on the former than the latter.

The reading time data show faster reading times in the passive than the active at the verb, and 3 regions within the by-phrase (determine, adjective, conjunction). This replicates the results of Exp. 1 and Exp. 2, with exception that in Exp. 2 some of these regions were only numerically faster in the passive. There was no significant or even marginal difference at the head of the by-phrase, contrary to Exp. 2. The fact that no region demonstrated longer reading times for the passive than the active contrasts with what is predicted by syntactic complexity-based, usage-based or heuristics-based accounts of passive sentence processing.

Regardless of predicate type, accuracy was lower in passives with respect to actives. No significant interaction between passivization and predicate type was found, contrary to what was predicted based on the results of Exp.1 and 2. The verification questions in Exp. 1 and 2 targeted theta-roles and other sentential aspects, while questions in Exp. 3 only targeted theta-role assignment, thus providing a more powerful test of interpretation difficulty. Thus, it may be that an effect on accuracy was not observed in Exp. 1 because the test was not sufficiently sensitive across an adequate number of trials (although this effect was clearly seen with stative predicates in Exp. 2, indicating some further variability).

There are two other possible factors that may have contributed to the disparity in accuracy across Exp. 1 and 2 vs. 3: (1) participants differences and (2) within-subject design effects. Both factors could contribute to the observed large participants variance in the difference in accuracy (between passives and actives) in Exp. 3 compared to 1 and 2, as we will explain. Despite random sampling, concern of a tertiary participant variable affecting the results is strengthened by the large variability across participants in the accuracy difference (passive-active) in Exp. 3. The question then becomes: what is varying across participants? The most studied inter-subject variability measure in sentence comprehension accuracy is working memory (WM) span. Indeed, WM has been found to correlate with various linguistic tasks (e.g., sentence comprehension: Daneman & Carpenter, 1980; self-paced reading: Kim & Christianson, 2013; Swets, Desmet, Hambrick & Ferreira, 2007) and various studies have shown that WM correlates with comprehension difficulty (e.g., Caplan et al., 2011; Just & Carpenter, 1992; Roberts & Gibson, 2002). It may be the case that passives are not difficult to interpret, which would explain the absence of an online difficulty effect. Rather, the verification task may be more difficult with a passive sentence because storing a noncanonical sentence is more taxing either in terms of the full representation or in manipulating the structure for purposes of the task. If this is the case, then one possibility is that a greater WM span would

reduce the difference in accuracy between actives and passives. Moreover, according to Ferreira (2003), participants with a lower WM span should rely on heuristics more often, as their algorithmic parsing should be more fragile, thus showing a greater accuracy difference between active and passives than participants with a higher WM span.

Another factor that could contribute to the variability observed in Exp. 3 arises from the within subject design. Exp. 3 had fewer items per condition (from 10 in Exp. 1, to 15 in Exp. 2, to only 7 in Exp. 3) given the within subject, Latin square design divided the same items over 4 conditions rather than 2 or 3. The fewer items/condition could have introduced a large amount of noise in the dichotomous data across participants.

Finally, a lack of interaction between predicate type and passivization in accuracy in Exp. 3 with respect to Exp. 1 and 2 might be attributed to both eventive/stative predicates being present in the same experiment. The eventive passive might have primed participants towards a verbal interpretation of the stative predicates, thus reducing the effect of the temporary ambiguity.

The primary aim of Experiment 4 was to determine whether variability in WM span correlates with the difference in accuracy between passives and actives. This could help in identifying the source of the variability in accuracy across experiments and more importantly a potential key to understanding the offline difficulty effect observed with passivization. We added two additional memory tasks to the self-paced reading one: (1) sentence reading span task and (2) n-back task. Both are considered to be a reliable measure of working memory (e.g., Conway et al. 2005; Daneman & Carpenter, 1980; Swets et al., 2007). To control for the possible presence of noise due to a small number of items per condition and for the priming of verbal interpretations, we used a between-subject design over predicate type. This doubled the number of items per condition while keeping everything else identical across experiments, thus

allowing for a between-subject analysis of a possible interaction between passivization and predicate type.

Experiment 4

Experiment 4 aimed at identifying whether the difference in accuracy across passive and active sentences could be related to individual differences in WM capacity.

A survey of the literature on the relationship between WM and language processing shows that the two most common measures of WM span are the *sentence span task* and the *n-back task* (e.g., Boyle, Lindell & Kidd 2013; Conway et al., 2005; Daneman & Carpenter, 1980; Dede et al., 2004; Jaeggi et al., 2010; Kim & Christianson, 2013; Roberts & Gibson, 2002; Sprouse et al., 2012; Swets et al. 2007). General agreement exists on the strong correlation between the reading span task (Daneman & Carpenter, 1980) and sentence comprehension accuracy (e.g., self-paced reading: Kim & Christianson, 2013; Swets et al., 2007). Likewise, studies show that the N-back task correlates with sentence comprehension accuracy (e.g., Conway et al., 2005). Nonetheless, these two tasks seem to be assessing different aspects of WM, as previous studies (Jaeggi, Buschkuehl, Jonides & Perrig, 2008; Jaeggi et al., 2010; Kane, 2005) found a low correlation between the n-back and the reading span task, suggesting that the two tasks tap into different working memory constructs. The reading span task requires participants to perform a cognitive task, e.g., judging whether a sentence is semantically plausible or not, and generating interference with the memory task, i.e., serial recall of letters. For this reason, the reading span task is believed to tap into active storage and processing functions necessary to perform linguistic tasks, such as sentence comprehension (Conway et al., 2005; Daneman & Carpenter, 1980). Additionally, the memory task performed during the reading span task requires serial recall of a list of items. On the contrary, the n-back task is largely based on externally triggered recognition. Processes

involved in the task include encoding new items and their position in a pre-existing list, comparison of the current item with the stored list, inhibition of response to irrelevant stimuli and finally updating information (e.g., stored list; Kane & Engle, 2002).

If (part of) the variance observed in the difference in accuracy across passives and actives was caused by inter-subject variability in WM span, we should find a correlation between the accuracy difference and the participants' memory span. That is, people with a lower WM span should have a larger difference in accuracy. This could indicate that people with a lower WM span rely more on the heuristic to process sentences, as predicted by Ferreira's model (2003; see also Christianson, et al., 2006, and Karimi & Ferreira, 2016) or that maintaining or manipulating the representation of a passive sentence for the comprehension task is more demanding. However, the consistently faster reading times observed for passives over actives seem to indicate that processing a passive step-wise is not difficult. Rather, maintaining/operating on the full representation of a passive sentence creates difficulty, and, possibly, even more so in participants with a lower WM span.

In order to investigate the contribution of WM to our accuracy effect, we used the reading span task and the n-back task. If our interpretations of Experiment 3 are correct and it provided a more reliable test of interpretation difficulty given that the comprehension questions always tested theta role assignment, then we should expect to see lower accuracy on passives than actives, again, in the current experiment. Moreover, if the lower accuracy on passives is related to greater demands on WM in completing the comprehension task, we would expect to find a correlation between the difference in accuracy (between passives and actives) and WM.

Finally, the new experiment should also test the replicability of the reading time findings from Exp. 1, 2 and 3. Replication is of significant value, given the observed replication crisis in Psychology (e.g., Ito, Martin & Nieuwland, 2016; Shanks et al., 2015). In terms of the

replication of reading time data, we expect faster reading times in the passive than the active sentence.

Methods and Design

1. Participants

A hundred and one native British English-speakers were recruited to participate in the study (68 females; average age: 24). A larger sample of participants was used in order to ensure sufficient variability in WM to test our hypothesis.

The same recruitment criteria and procedure as in Exp. 1-3 was used. None of the participants had previously participated in any of the experiments related to this project (including the pre-norming tests).

2. Stimuli

As mentioned in the Discussion section of Exp. 3, to control for the possible presence of noise due to a small number of items per condition, we split the items of Exp. 3 in a between-subject design, according to predicate type while syntax of the sentence (active vs. passive) remained a within-subject factor.

The experiment contained 2 sentence sets, each with 28 items taken from Exp. 3. Everything else (types of filler sentences; overall sentence length) was also identical to Exp. 3.

3. Memory tasks.

Reading span task. The reading span task consisted of 2 separate, but intermixed tasks. The first task was a sentence judgment task: participants were asked to decide whether the sentences (10 to 15 words long) presented to them on a PC screen were correct or incorrect. As specified in the instructions, their decision had to be based on semantic/pragmatic, rather than grammatical, considerations (see Table 7 for example sentences). Each item was presented at

the centre of the screen and would automatically disappear after 1000ms, regardless of whether the participant had made a decision or not. Participants were instructed to be as accurate as possible (i.e., maintain an overall accuracy in the judgment task higher than 85%) and received feedback on their accuracy after each trial. [Table 7 near here]

The second task consisted in a letter recall task. After each judgment, a letter (only uppercase consonants: F, H, J, K, L, N, P, Q, R, S, T, Y) would appear on the screen for 1000ms. At the end of each block of trials, participants were asked to recall all the letters in the correct order and received feedback on their performance (see Figure 4). Each block could contain between 2 and 5 sentences.

Participants completed 6 practice trials of the first task, 3 practice trials of the second task and 3 practice trials of the two tasks interleaved. Each trial contained 2 to 5 items (i.e., sentences plus letter) that were presented in a random order, rather than ascending order, to avoid the possibility that participants would rely on strategies (e.g., proactive interference) that might come from anticipating the number of items to recall per trial. Each set of items (2,3,4,5) was presented 3 times, hence participants completed 12 trials. Feedback was given on both tasks to ensure participants' engagement, which implied that both the storage and processing functions of working memory were actually at work. [Figure 4 about here]

n-back task. In the n-back task participants were presented with a series of letters (only uppercase consonants) on a PC screen and had to respond (by pressing the "A" key on the keyboard) when the letter currently presented was identical to the one presented n-positions back (e.g., 1-, 2-, 3-positions back; see Figure 5 for an example). Participants received 9 trials of practice per level tested, and were tested on 3 blocks of 15 trials (i.e., letters) per level. In each block, 5 trials were targets and 10 were not. Each letter was presented for 500ms. When the letter disappeared, a blank screen appeared for 2000ms. Hence, each trial lasted 2500ms. [Figure 5 about here]

2. Procedure

For the self-paced reading task, the same procedure as in Exp. 3 was used.

In order to maintain the same procedure as in Exp. 1, 2 and 3, the self-paced reading task was always administered first. To avoid possible order effects, the 2 memory tasks were then presented in a counterbalanced order, i.e., half of the participants performed the reading span task first, and half the n-back task first.

3. Data analysis

In the self-paced reading task, 5 participants scored lower than the exclusion threshold in accuracy on fillers (75%), hence they were excluded from the final analysis. Data from 96 participants were analysed.

The analysis of reading time, accuracy and Reaction Times data followed the same steps as in Exp.1, 2 and 3. 0% of the original response time data were removed.

Correlation analysis. Residual logRTs, logRTs and accuracy data were further analysed with respect to WM measures. The following measures were inserted, together with either the reading span or the n-back span measure, in a correlation analysis: difference in accuracy between active and passive, difference in Reaction Times between active and passive, difference in Reading Times (entire sentence and critical regions only) between active and passive, mean overall accuracy, mean accuracy on actives, mean accuracy on passives, mean overall Reaction Times, mean Reaction Times to actives, mean Reaction Times to passives, mean overall Reading Times, mean Reading Times for actives, mean Reading Times for passives.

Reading Span Task. Data from the reading span task were scored using the following procedure. One point was given to each trial if the subject had responded correctly to both

tasks, i.e., if the participant had correctly judged the sentence and correctly recalled the letter in its serial position. Hence, each participant obtained an absolute score out of 42 (total number of trials; Kim & Christianson, 2013; Swets, et al. 2007). This score was then standardized to avoid collinearity (Belsey, 1991).

n-back Task. Data from the n-back task were scored by subtracting the total number of false hits across blocks (when the participant pressed "A" even if the letter currently presented was not a target) from the total number of correct hits across blocks. The score was then divided by the total number of blocks (Jaeggi et al., 2010) and the results standardized to avoid collinearity (Belsey, 1991).

The standardized scores of the reading span task and the n-back task were then inserted in the fixed effects structure of both the linear mixed effects model used to analyse reaction and response times of the self-paced reading task, the mixed effects logistic regression used to analyse accuracy data collected in the self-paced reading task and our correlation analyses. The analyses were performed in R (https://www.rstudio.com/).

Results

Comprehension Question Results:

Offline accuracy and response time results replicated the results of Exp. 2 and 3. There was a significant difference in comprehension accuracy following a passive with respect to an active sentence (active: 85.42%; passive: 79.76%; β=-.59, z=-3.38, p<.001[10]; see Table 8). No other effect of accuracy was significant.

Response times to comprehension questions following passive sentences were significantly longer than following their active counterparts (β=.07, t=6.2, p<.001; see Table 8), similarly to what we found in our previous experiments. Moreover, we found a significant

effect of memory span as measured by the sentence span task (β=-.03, t=-2.79; p=.005), indicating that people with a larger memory span were faster in answering comprehension questions. No other effect was significant. [Table 8 about here]

In order to exclude possible learning effects that might have confounded our results, we ran and compared 2 separate analyses on the data collected in the first vs. second half of the sessions[12]. The analysis did not reveal any significant difference with respect to the overall results.

Reading Time Results:

The analysis of the reading times for correct trials only replicated the results of our previous experiments, but more robustly. Passives were read significantly faster than actives at the verb (β=-.03, t=-3.47, p<.001), at the determiner of the by-phrase (β=-.05, t=-8.79, p<.001), at the first adjective (β=-.02, t=-3.16, p=.002), at the conjunction (β=-.03, t=-3.78, p<.001), and at the second adjective[11] (β=-.02, t=-2.6, p=.009, see Figure 6). However, as reported in Exp. 2, this trend reversed at the head of the by-phrase, where passive sentences were read marginally slower than active sentences (β=1.84, t=1.89, p=.059), signalling a possible point of revision or difficulty in integration.

No effect of predicate type on reading time data was found.

An effect of WM span as measured by the n-back task was only found at the conjunction (β=-.01, t=-2.1, p=.04), indicating that people with a larger memory span were faster at reading this region. No other effect was significant. [Figure 6 about here]

Finally, to examine possible learning effects, we ran and compared 2 separate analyses on the data collected in the first vs. second half of the sessions. The analysis did not reveal any significant difference with respect to the overall results.

WM Correlations:

With respect to WM measures, our correlation analyses revealed a significant correlation between Response Times to comprehension questions and WM span as measured by both the sentence span task (r=-.22, t=-2.19, p=.03) and the n-back task (r=-.23, t=-2.28, p=.03). The greater the WM span the faster the response time. We found a correlation between the difference in accuracy in active vs. passive sentences and WM span as measured by the sentence span task (r=-.2, t=-2.01, p=.047), due to the fact that participants with a lower WM presented a larger difference between active and passive sentences. Moreover, WM span as measured by the sentence span task correlated with accuracy on passive sentences (r=.22, t=2.2; p=.03), due to the fact that participants with a lower WM span performed worse on passives, but not active sentences, or accuracy overall. The n-back task scores did not correlate with any accuracy measure (average accuracy, difference in accuracy between actives and passives, accuracy on actives only, accuracy on passives only). Finally, there was no correlation between online data and WM span. Hence, both the sentence reading span and the n-back span appear to be good predictors of offline rather than online processing. These results are consistent with what is reported in the literature on working memory, which found effects of memory span on offline rather than online processing (Caplan et al., 2011; Evans et al., 2014).

Overall, the data confirm that passive sentences are processed faster than active sentences, up to the head of the by-phrase where participants are marginally slower in passives with respect to actives.

Discussion

The main aim of this experiment was to determine whether WM could account for the offline difficulty effect observed in accuracy. Additionally, it investigated whether the within-subject design of Exp. 3 limited its ability to detect an interaction between passivization and predicate type. Just as in Exp. 3, we found passives were understood less accurately than their active

counterparts, and no interaction between passivization and predicate was observed. As suggested in the discussion of Exp. 3, Exp. 1 and 2 may not have been sufficiently sensitive to show as reliable an effect of passivization on accuracy, as not all questions targeted theta-roles. Interestingly, WM as measured by the sentence span task correlated with the difference in accuracy between actives and passives as well as accuracy on passives only. However, there were no effect of sentence span scores or interaction between accuracy and sentence span scores in our regression analysis, indicating that a difference in WM span cannot fully account for the difference in accuracy between passives and actives. Moreover, we found that WM span significantly correlated with reaction times to comprehension questions overall (a significant effect of sentence span also emerged in our linear mixed effects model). However, WM span generally did not correlate with reading time data (with exception of the conjunction). Again, reading time data showed passives were processed faster than actives up to the head of the by-phrase, where the pattern reversed, and passives were read marginally slower than actives, very similar to results from Exp. 1-3.

Nonetheless, the correlation between WM and both the difference in accuracy across the voice manipulation and the accuracy on passive sentences only, suggests that the offline complexity in accuracy is partially due to variation in WM span. This offers some explanation for variability across studies in terms of this difference. It is possible that even with random sampling there was some biases in the WM spans of the participants across experiments. WM span was also found to be a good predictor of participants' speed and ability to accurately perform the comprehension task. The better their WM span the more accurate they were to match the meaning of the question to that of the preceding sentence and did so more quickly. The lack of correlation between WM and the online data is problematic for a unified perspective of passive sentence processing across online and offline tasks. The contrast

between online reading times and offline accuracy in terms of their relation to WM, further confirms that the offline and online tasks stem from (at least partially) independent processes.

While the accuracy data are compatible with heuristic-based (Ferreira, 2003), syntactic-based and usage-based models of passive sentence processing, which predict passive sentences to be more difficult to process than active sentences, the reading time data are not. Rather than longer reading times with passives sentences, we see shorter ones. At present, the only processing model compatible with our reading time facilitation for passives is an expectation-/surprisal-based account of parsing (e.g., Hale, 2001; Levy, 2007). Why passives show lower accuracy in comprehension is partially explained by their greater demands on WM for the task. Again, a speed-accuracy trade-off cannot offer a good explanation, as the faster reading times were based on accurate trials alone.

As observed in our previous studies, online and offline data offer a contrasting picture of passive sentences processing. These diverging results are mirrored by an overall independence of accuracy and reading time measures as they differentially relate to WM span. WM measures correlated with our offline but not online measures. A possible solution to this puzzle will be further considered in the General Discussion section.

General discussion

Across four self-paced reading experiments, two reliable, seemingly contradictory, results emerged. On the one hand, we observed faster reading times with passives than actives at the verb and multiple regions of the by-phrase. On the other hand, we observed that passives were comprehended less accurately than actives. While the reading time difference between actives and passives reversed direction at the post-verbal noun, it never reached significance in any of the four experiments. Lastly, despite our predictions, we did not find that passivization interacted with predicate type in either the reading time or accuracy data. Although there was

some indication of an interaction in accuracy across Exp. 1 and 2, Exp. 3 and 4 found no such tendency when comprehension questions consistently addressed thematic role assignment, the most difficult aspect of passive interpretation.

The three main accounts for passive difficulty are not commensurate with the two main findings. Likewise, a simple speed-accuracy trade-off account of these two effects is not adequate, as the reading time facilitation was found even when analysing accurate trials alone. It is possible that there was an online-speed, offline-accuracy trade-off, whereby online readers sped-up in the passive/progressive (for unknown reasons) such that offline further integration or interpretational effort was required. This would then explain the greater number of errors offline and longer response times in these conditions[1]. However, in such a speed-accuracy trade off account you would not expect the reading speed to be systematically modulated as we observe. The faster reading times corresponded to points following the additional morphology in these conditions which is less compatible with a generic increase in speed (online) at the cost of accuracy (offline) but, as we will argue below with further supporting data, more compatible with a surprisal account. Also, any account that resorts to frequency of the verbs in the passive vs. active voice is inadequate given the results from our BNC analysis. No single account is compatible with both results. We must turn away from the main-stream views of passive processing and consider independent explanations for each of these results, which raises two questions: (1) why do passives facilitate reading, but (2) make comprehension difficult? Our answers will leave us with the conclusion that passives are not difficult to parse and interpret. Rather, the offline difficulty with passives arises from particular tasks that require greater processing demands on memory.

Passives Facilitate Reading Times:

[1] Statistically testing this account is complicated by the overall low number of trials with errors compared to the number of trials with correct responses.

The faster reading times of passives is compatible with surprisal or expectation-based models (e.g., Hale, 2001; Levy, 2007). These models predict passives to be processed faster than actives in regions where morphological cues (i.e., auxiliary, past participle, "by") provide expectations about upcoming syntactic categories. Indeed, the facilitation (i.e., faster reading times) was most robust at the regions immediately following the additional morphology (e.g., verb, determiner). Experiment 1 further supports this conclusion by demonstrating a distinction between progressives and perfective actives at the verb. The progressive which has an additional morphological cue (i.e., auxiliary) than the perfective had shorter reading durations at the verb. A search of the Brown corpus supports this interpretation. The probability of a verb following an initial NP and then the auxiliary "was" is 0.92, whereas the probability of a verb following an initial NP alone is 0.66. Further work should establish whether the critical factor is the number of potential categories that can occur subsequently or the simple probability of that particular category. As these two factors are correlated, it is not currently transparent which is most critical.

Likewise, the determiner of the by-phrase/object was read faster in the passive than the active. The presence of the "by" would again provide further expectation for a determiner to follow with respect to a verb in isolation. In some experiments, the facilitation had a roll-on effect and emerged in subsequent regions (i.e., adjective and coordinator).

Another potential explanation for the difference in reading times between passives and actives in the region following the verb (by-phrase vs. internal argument) focuses on the adjunct vs. argument status of this region (Tutunjian & Boland, 2008). The by-phrase is argued to be more adjunct-like than an internal argument, because it can be omitted without affecting grammaticality (e.g., Grimshaw, 1990). However, the by-phrase satisfies a thematic role of the predicate and is syntactically active even when omitted[13], demonstrating argument-like properties. An eye-tracking while reading study found first pass reading times were shorter for

argument PPs than adjunct PPs in temporarily ambiguous structures. It was argued that the facilitation in processing arguments was due to their projection already at the verb[14] (Boland & Blodgett, 2006). The direction of the observed reading time difference in our studies is inconsistent with this: the "adjunct" condition (i.e., passive) is read faster than the "argument" one (i.e., active). The faster reading times of both passives and argument PPs (Boland & Blodgett, 2006) may collectively be captured by a surprisal-based account.

While the only reliable reading time result was that passives were read faster, we saw a marginal effect at the post-verbal noun which was read slower in the passive than the active in 2 experiments. Given it did not even approach significance (p-value>.7) in the other two experiments, this result does not warrant significant discussion. However, it is worth noting that if passives are complex, they are only weakly so (online) and not comparable to complexity effects observed in other cases of non-canonical structures. For example, object vs. subject relatives/clefts show clear complexity effects both online and offline (Garnham & Oakhill, 1987; Ferreira, 2003). Further, even if the effect were reliable/significant, it would not be compatible with the location where the agent-first heuristic or syntactic accounts would predict the complexity to occur (verb).

On the contrary, an effect at the post-verbal noun could be consistent with an interference effect in integrating the by-phrase with the VP, whereby the VP has to be merged lower than the by-phrase, unlike in actives where the object linearly merges with the VP. The location at which this apparent effect arises, i.e., the head of the by-phrase, also speaks in favour of a "structural" account, rather than a more simplistic "delay" account. Indeed, this is commensurate with more recent theoretical analyses of passives (Alexiadou, Anagnostopoulou & Schäfer, 2018; Collins, 2005; Gehrke & Grillo, 2009).

While we argue that the current online data are not compatible with *Good Enough* processing when it comes to an agent-first heuristic, we are not arguing against *Good Enough*

processing under all conditions. As our goal was to assess why passivization is difficult, we focused on the agent-first heuristics and not other possible ones, e.g., a semantic heuristic that language users apply in sentence processing as suggested by various other data (Christianson, Luke & Ferreira, 2010).

In all of our readings of *Good Enough* parsing the heuristic applies prior to completion of algorithmic processing. So long as this is the case, our online data are problematic for this account when it comes to passivization complexity. It could, however, be argued that self-paced reading is not ideal for detecting a "revision" effect. Self-paced reading does not distinguish between initial (i.e., first-pass) processing and later revisions (i.e., re-reading). However, it is very likely that revision processes contribute to the reading times measured with SPR. The language comprehender must decide between holding off on any revision to a later point while increasing storage demands vs. engaging in revision and reducing storage demands before viewing the next word. It is unlikely that the comprehender will always increase storage demands and so reanalysis should contribute to reading times. The fact that we see *shorter* reading times at the verb and into the by-phrase does not fare well for reanalysis. Again, it is worth remembering that other noncanonical sentences do show longer reading times in SPR at the verb. This indicates that noncanonical structures are not processed equivalently; a finding that is problematic for both heuristics and syntactic complexity. Future studies should use eye-tracking with measures that more directly relate to "reanalysis" to further support these conclusions.

While our readings of the literature suggest that the heuristic applies prior to algorithmic processing, a reviewer (Kiel Christianson) points out that the heuristic can apply after the sentence. If so, then it would remain compatible with our reading time data. This however would seem more commensurate with a task strategy rather than an approach to language

comprehension. We will elaborate on a possible form of strategy that could account for the data later in the discussion.

While in the above we argue that there is no online difficulty effect, we want to emphasize that this is not evidence against movement itself. It may be that movement is only difficult for the parser in particular structural configurations. Similarity-based interference in the case of object vs. subject relative clauses, for instance, shows that movement is difficult when the dependency crosses a "similar" constituent to the antecedent (Fedorenko, Gibson & Rohde, 2006; Gordon et al., 2001; Van Dyke & McElree, 2006).

Passives Impair Comprehension:

Unlike the reading time data, the accuracy data implies that passives are harder to interpret than actives. How can we reconcile a difficulty effect in interpretation with one of facilitation in reading time?

It seems plausible that the full interpretation of a passive interacts with the task in generating these offline difficulty effects. Our comprehension task involved retrieving/maintaining the experimental sentence's interpretation to confirm whether it matches the interpretation of the comprehension question. Participants may rely to some degree on the surface form for completing the task. In doing so, passives may be more susceptible to memory interference between the experimental and comprehension question. Supporting evidence for this interpretation comes from our finding that WM correlated with the difference in accuracy between actives and passives. However, this effect was only observed in our correlation analysis and not in our regression model, suggesting that WM cannot solely account for the difference in accuracy between active and passive sentences.

While our measures of WM focused on a verbal store, what may be taxing to store or manipulate in the case of passives may be a semantic representation. These demands may not be fully assessed by our WM test. Participants may encode the sentence as a simplified active-

like representation (Agent-Verb-Patient, e.g., dog bite man) for the purpose of the task (see Anderson, 1974, for a very similar perspective). This "active-like" representation may be more robust to memory decay particularly when the task is focused on meaning rather than form. However, generation of this representation would be more difficult for the passive than the active structure. A passive sentence would require an additional transformation, i.e., a "reverse" movement operation, to generate a post-interpretive representation. Hence this operation could be more susceptible to errors or may result in falling back on the less reliable form-based representation, on some trials (which would also lead to more errors and explain our correlation effects). This perspective argues that deriving the passive representation and interpretation is not taxing, step-wise, but operating over the entire (passive) interpretation in memory is. Bear in mind that the difference in accuracy between active and passive lies in the 5% range, hence they are often processed effectively and accurately. While movement/non-canonicity may not be complex to parse and interpret in certain structures (i.e., passives), maintaining or working with such a representation for the task may be.

Similar to the argument just proposed for the current comprehension task, it can be argued that the task used by Ferreira (2003) would be more difficult for the passive condition than the active one. Recall from the Introduction section that the task involves distinguishing *doer* from *acted on*. In the case of a passive the participant needs to maintain a distinction between "subject" and "agent". In the case of an active, the two coincide, making the task less taxing and susceptible to interference/confusion.

An interesting avenue for future work is to look at additional factors, beyond WM, that might explain the variability in the accuracy difference between active and passive sentences. Previous work has found level of academic achievement to be a factor in interpretive performance on passives (Dąbrowska & Street, 2006; Street & Dąbrowska, 2010). These studies investigated differences between those enrolled in postgraduate degrees vs. non-

graduates. Given that we sampled from the UCL University pool, the effects of academic achievement should be rather limited in our studies.

Overall, these conclusions point to further work in providing better linking hypotheses between parsing and interpretation and post-interpretive processes associated to the task.

Predicate Type:

Finally, predicate type did not interact with passivization. This is in contrast with our expectations given that predicate type interacts with passivization for availability and interpretation (Belletti & Rizzi, 1988; Gehrke & Grillo, 2009; Snyder & Hyams, 2015). Given the constraints that statives face when passivized, we expected them to be more difficult to process when presented in the passive form. Two possible explanations for our null result can be provided. Firstly, it is possible that the theoretical difference reported cross-linguistically (Belletti & Rizzi, 1988; Gehrke & Grillo, 2009; Snyder & Hyams, 2015) has no psycholinguistic counterpart: more constraints on use do not necessarily determine processing complexity. This would contrast with some results from aphasia and acquisition (Maratsos et al., 1985; Grodzinsky, 1995; Volpato et al., 2013) and from recent results collected by Ambridge and colleagues (2016). However, it is consistent with the conclusion drawn from Messenger et al. (2012): the interaction between predicate type and passivization is only observed with particular tasks (i.e., sentence-to-picture matching). Messenger et al. (2012) argue that the difficulty children have with passives of stative(-like) predicates compared to eventive ones emerges in sentence-to-picture matching tasks because of their difficulty encoding such predicates pictorially. Children demonstrate syntactic priming for stative passives, just as they do with eventive passives indicating they are equally-well acquired. This argument is in many respects in the same vein as the one we have presented for the overall difficulty of passivization, independent of predicate type.

In contrast to the claims by Messenger et al (2012), a recent study by Ambridge et al (2016) with healthy adults demonstrates an interaction between semantics of the predicate and passivization in both acceptability judgments and response times in a video verification task. Ambridge et al. (2016) investigated semantic affectedness rather than states vs. events, as in our study. However, the predicate with the least semantic affectedness (i.e., on the internal argument) corresponded to subject experiencers (our stative predicate) and those with the most affectedness corresponded to agent-patient verbs (our eventive predicates). The interaction is in the direction we predicted: greater difficulty with the stative-like passives than eventive-like passives. The semantic richness of our sentences (referential NPs and for-clauses) may have helped to support the eventive reading of the stative predicate. Alternatively, the interaction in Ambridge et al. (2016) may be due to verb selection criteria. In Ambridge et al. (2016) the interaction observed in study 2 appears to stem *solely* from the non-passivizable predicates (e.g., belch). In the third experiment, the non-passivizable predicates are excluded as are a large proportion of predicates from each verb class (92% of agent-patient, 56% of experiencer-theme). This raises the question as to whether the interaction is an artefact of predicate selection. In combination with the Messenger et al. (2012) data, it is necessary to further understand which tasks and stimulus parameters give rise to this interaction to better understand its nature.

Conclusions

The comprehension of passive sentences has generally been considered more difficult than that of active sentences (Bever, 1970; Ferreira, 2003). However, previous *online* studies do not conclusively support this general assumption.

The present studies aimed at investigating the puzzling results from previous literature by directly comparing online and offline processing of passive sentences while additionally

controlling for the predicate semantics. Experiment 1 found eventive passives to be processed faster than actives online, and resulted in equally good performance as actives offline. Experiment 2 found significant evidence of offline difficulty for stative passives, and a trend towards a greater difficulty in online processing of passives at the head of the by-phrase. Experiment 3 found passives to be processed faster than actives online, but harder to comprehend offline, regardless of predicate type. Most likely, Experiment 3's use of theta-role questions alone provided a more sensitive test of interpretation difficulty than Experiment 1 and 2. Experiment 4 replicated the results of Experiment 3, and additionally found WM measures correlated with the difference in accuracy between actives and passives, but not with the reading time data. Finally, across experiments we found no interaction with predicate type, either online or offline, contrary to what might be expected based on the theoretical literature and work in acquisition, aphasia and healthy adults (Belletti & Rizzi, 1988; Gehrke & Grillo, 2009; Snyder & Hyams, 2015; Ambridge et al., 2016).

Overall, the data collected in the present experiments contrast with mainstream theories of passive sentence processing in showing that they are not more difficult to process than actives (Chomsky 1981; Ferreira, 2003; John & Jones, 2015; Kiparsky, 2013). Rather, the reading time data are compatible with expectation-based (e.g., Levy, 2007) or surprisal-based accounts (e.g., Hale, 2001). This interpretation might at first appear at odds with the offline finding that passives are more errorful than actives in interpretation. However, we argue that these effects arise from task-related demands. In particular, the whole passive interpretation is difficult to maintain in a form robust to memory decay/interference. The correlation between WM and the difference in accuracy across voice supports this.

Future work needs to further investigate these interpretations by considering additional online and offline measures and directly comparing different populations with the same manipulations and experimental procedures. Moreover, these results critically draw attention

to the need for greater theoretical analysis of the link between offline tasks and online processing.

Acknowledgments

This research was partly funded by the DFG -- Leibniz Prize AL 554/8-1 awarded to Artemis Alexiadou. We gratefully acknowledge the DFG contribution. We would also like to gratefully thank John Hale for assistance with the Brown Corpus search.

[1] We only consider the plausible passives (either arguments could be agent or patient) given our interest is the complexity of passive syntax/word order. However, a similar difference was found with the implausible conditions.

[2] Model did not converge with random slopes. The same model was used for the analysis of theta-role questions only and for first- vs. second-half analysis, given that more complex models did not converge.

[3] Model only converged with random slope for subjects but not for items. The same model was used for the analysis of theta-role questions only and for first- vs. second-half analysis, given that more complex models did not converge.

[4] Data from trials 7-51 (the first 6 items were practice items) were analysed as "first-half" and data from trials 51-96 as "second-half".

[5] The data were divided in first vs. second half as per Experiment 1.

[6] For all the reading times analyses, the most complex model, including both intercept and random slope for both subjects and items, always converged. The only exception was the first adjective region, where the model only converged with intercept for both subjects and items and slope only for subjects and not items.

[7] Model only converged with intercepts.

[8] Data from trials 7-51 (the first 6 items were practice items) were analysed as "first-half" and data from trials 51-96 as "second-half".

[9] Model for the verb region analysis only included Syntax in the structure of the random effect of the items, as the verb differed across predicate type.

[10] For all accuracy analyses, the model only converged with random intercepts and not slopes.

[11] The data were divided in first vs. second half as per Experiment 1, 2 and 3.

[12] The model only converged with syntax and not predicate type in the structure of the Item random effect.

[13] The presence of a silent argument in short passives is supported by several syntactic diagnostics, including the ability to support subject controlled infinitival sentences and subject-oriented modifiers and depictives (e.g., "The book was written to collect the money/deliberately/drunk"; Baker, 1988;

Manzini, 1983; Roeper, 1987), and to bind reflexives (e.g., “such privileges should be kept to oneself”; Baker, Johnson & Roberts, 1989; Roberts, 1987).

[14] This perspective seems as though it could also be accounted for under a surprisal based account.

References

Alexiadou, A., Anagnostopoulou, E., & Schäfer, F. (2018). Passive. In N. Hornstein, H. Lasnik, P. Patel-Grosz, Pritty & Ch. Yang (Eds.), *Syntactic Structures 60 Years On. The Impact of the Chomskyan Revolution in Linguistics*. Berlin: Mouton DeGruyter.

Ambridge, B., Bidgood, A., Pine, J.M., Rowland, C.F., & Freudenthal, D. (2016). Is Passive Syntax Semantically Constrained? Evidence From Adult Grammaticality Judgment and Comprehension Studies. *Cognitive Science*, *40*, 1435–1459. doi: 10.1111/cogs.12277

Anderson, J.R (1974). Verbatim and Propositional Representation of Sentences in Immediate and Long-Term Memory. *Journal of Verbal Learning and Verbal Behavior, 13*, 149-162. doi: 10.1016/S0022-5371(74)80039-3

Barr, D.J., Levy, R., Scheepers, C., & Tily, H.J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255-278. doi: 10.1016/j.jml.2012.11.001

Belletti, A. & Rizzi, L. (1988). Psych Verbs and Theta Theory. *Natural Language and Linguistic Theory*, *6*, 291-352. doi: 10.1007/BF00133902

Belsley, D. (1991). *Conditioning Diagnostics: Collinearity and Weak Data in Regression*. New York: Wiley.

Bever, T.G. (1970). The cognitive basis for linguistic structures. In J.R. Hayes (Ed.), *Cognition and the development of language* (pp. 279–352). New York: Wiley

Boland, J.E., & Blodgett, A. (2006). Argument Status and PP-Attachment. *Journal of Psycholinguistic Research*, *35,* 385–403. doi: 10.1007/s10936-006-9021-z

Borer, H., & Wexler, K. (1987). The maturation of syntax. In T. Roeper & E. Willliams (Eds.), *Parameter setting and language acquisition* (pp. 123–172). Dordrecht: Reidel.

Boyle, W., Lindell, A. K., & Kidd, E. (2013). Investigating the Role of Verbal Working Memory in Young Children's Sentence Comprehension. *Language Learning*, *63*, 211–242. doi: 10.1111/lang.12003

Caplan, D., DeDe, G., Waters, G., Michaud, J., Tripodis, Y. (2011). Effects of age, speed of processing, and working memory on comprehension of sentences with relative clauses. *Psychology and Aging*, *26*, 439-450. 10.1037/a0021837

Caplan, D., Vijayan, S., Kuperberg, G., West, C., Waters, G., Greve, D., & Dale, A.M. (2002). Vascular responses to syntactic processing: event-related fMRI study of relative clauses. *Human Brain Mapping*, *15*, 26-38. doi: 10.1002/hbm.1059

Caplan, D., & Waters, G.S., (2013). Memory mechanisms supporting syntactic comprehension. *Psychonomic Bulletin & Review*, *20*, 243-268. doi: 10.3758/s13423-012-0369-9

Caplan, D., Waters, G., Dede, G., Michaud, J., & Reddy, A. (2007). A study of syntactic processing in aphasia I: behavioral (psycholinguistic) aspects. *Brain and Language*, *101*, 103-50. doi: 10.1016/j.bandl.2006.06.225

Carrithers, C. (1989). Syntactic complexity does not necessarily make sentences harder to understand. *Journal of Psycholinguistic Research*, *18*, 75-88. doi: 10.1007/BF01069048

Chomsky, N. (1981). *Lectures on Government and Binding.* Foris, Dordrecht.

Chow, W., Momma, S., Smith, C., Lau, E., & Phillips, C. (2016). Prediction as memory retrieval: timing and mechanisms. *Language, Cognition and Neuroscience*, *31*, 617-627. doi: 10.1080/23273798.2016.1160135

Christianson, K. (2016). When language comprehension goes wrong for the right reasons: Good Enough, underspecified, or shallow language processing. *Quarterly Journal of Experimental Psychology*, *69*, 817-828. doi: 10.1080/17470218.2015.1134603

Christianson, K., Hollingworth, A., Halliwell, J.F., Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology*, *42*, 368–407. doi: 10.1006/cogp.2001.0752

Christianson, K., Luke, S. G., & Ferreira, F. (2010). Effects of plausibility on structural priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 538–544. doi:10.1037/a0018027

Christianson, K., Williams, C. C., Zacks, R. T., & Ferreira, F. (2006). *Misinterpretations of garden-path sentences by older and younger adults*. Discourse Processes, *42*, 205-238. doi:10.1207/s15326950dp4202_6

Collins, C. (2005). A Smuggling approach to the passive in English. *Linguistic Inquiry, 36*, 289-297. doi: 10.1111/j.1467-9612.2005.00076.x

Conway, A.R.A., Kane, M.J., Bunting, M.F., Hambrick, D.Z., Wilhelm, O., & Engle, R.W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*, 769–786. doi: 10.3758/BF03196772

Dąbrowska, E., & Street, J. (2006). Individual differences in language attainment: Comprehension of passive sentences by native and non-native English speakers. *Language Sciences*, *28*, 604-615. doi: 10.1016/j.langsci.2005.11.014

Daneman, M., & Carpenter, P.A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning & Verbal Behavior*, *19*, 450-466. doi: 10.1016/S0022-5371(80)90312-6

Dede, G., Caplan, D., Kemtes, K., & Waters, G. (2004). The Relationship between Age, Verbal Working Memory, and Language Comprehension. *Psychology and Aging*, *19*, 601-616. doi: 10.1037/0882-7974.19.4.601

Dickey, M.W., & Thompson, C.K. (2009). Automatic processing of wh- and NP-movement in agrammatic aphasia: Evidence from eyetracking. *Journal of Neurolinguistics*, *22*, 563–583. doi: 10.1016/j.jneuroling.2009.06.004

Evans, W.S., Caplan, D., Ostrowski, A., Michaud, J., Guarino, A.J., & Waters, G. (2014). Working memory and the revision of syntactic and discourse ambiguities. *Canadian Journal of Experimental Psychology*, *69*, 136-155. doi: 10.1037/cep0000037

Fedorenko, E., Gibson, E., & Rohde, D. (2006). The nature of working memory capacity in sentence comprehension: Evidence against domain-specific working memory resources. *Journal of Memory and Language, 54*, 541-553. doi: 10.1016/j.jml.2005.12.006

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, *47*, 164–203. doi: 10.1016/S0010-0285(03)00005-7

Ferreira, F. & Christianson, K. (2016). Is Now-or-Never Language Processing Good Enough? *Behavioral and Brain Sciences*, *39*, 1-72. doi:10.1017/S0140525X1500031X

Ferreira, F. & Patson, ND. (2007). The 'Good Enough' approach to language comprehension. *Language and Linguistics Compass*, 1-2, 71-83. doi: 10.1111/j.1749-818X.2007.00007.x

Garnham, A., & Oakhill, J. (1987). Interpreting elliptical verb phrases. *Quarterly Journal of Experimental Psychology, 39*, 611–625. doi: 10.1080/14640748708401805

Gehrke, B., &, Grillo, N. (2007). Aspects on passives. *Proceedings of ConSOLE XIV*, 121-141. Retrieved from http://parles.upf.edu/llocs/bgehrke/home/console05_passives_ho.pdf

Gehrke, B., &, Grillo, N. (2009). How to become passive. In Grohmann & K. Kleanthes (Eds.), *Explorations of Phase Theory: Features and Arguments* (pp. 231-268). Berlin: Mouton de Gruyter.

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition*, *68,* 1-76. doi: 10.1016/S0010-0277(98)00034-1

Gordon, P., & Chafetz, J. (1990). Verb-based versus class-based accounts of actionality effects in children's comprehension of passives. *Cognition, 36*, 227-54. doi: 10.1016/0010-0277(90)90058-R.

Gordon, P.C., Hendrick, R., & Johnson, M. (2001). Memory interference during language processing. *Journal of Experimental Psychology*, *27*, 1-13. doi: 10.1037/0278-7393.27.6.1411

Gordon, P.C., Hendrick, R., Johnson, M., & Lee, Y. (2006). Similarity-based interference during language comprehension: Evidence from eye tracking during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 1304-21. doi: 10.1037/0278-7393.32.6.1304

Gordon, P.C., Hendrick, R., & Levine, W.H. (2002). Memory-load interference in syntactic processing. *Psychological Science*,*13*, 425-30. doi: 10.1111/1467-9280.00475

Grillo, A. (2008) *Generalized minimality: Syntactic underspecification in Broca's aphasia*. Doctoral dissertation distributed by LOT, University of Utrecht.

Grimshaw, J. (1990). *Argument Structure*. Cambridge, Mass.: MIT Press.

Grodner, D.J., & Gibson, E.A.F. (2005). Consequences of the Serial Nature of Linguistic Input for Sentential Complexity. *Cognitive Science*, *29*, 261–291. doi: 10.1207/s15516709cog0000_7

Grodzinsky, Y. (1990). *Theoretical Perspectives on Language Deficits*. Cambridge, MA: MIT Press.

Grodzinsky, Y. (1995). Trace deletion, theta-roles, and cognitive strategies. *Brain & Language*, *51*, 467-497. doi: 10.1006/brln.1995.1072

Grodzinsky, Y. (2000). The neurology of syntax: language use without Broca's area. *Behavioral and Brain Sciences*, *23*, 1-21. doi: 10.1017/S0140525X00002399

Hakes, B., Evans, J., & Brannon, L. (1976). Understanding sentences with relative clauses. *Memory and Cognition, 4*, 283–296. doi: 10.3758/BF03213177

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, 159–166. doi: 10.3115/1073336.1073357

Hofmeister, P. (2011). Representational complexity and memory retrieval in language comprehension. *Language and Cognitive Processes, 26*, 376-405. doi: 10.1080/01690965.2010.492642

Ito, A., Martin, A.E., & Nieuwland, M.S. (2016). How robust are prediction effects in language comprehension? Failure to replicate article-elicited N400 effects. *Language, Cognition and Neuroscience*, 1-12. doi: 10.1080/23273798.2016.1242761

Jaeggi, S.M., Buschkuehl, M., Jonides, J., & Perrig, W.J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 6829-6833. doi: 10.1073/pnas.0801268105

Jaeggi, S.M., Buschkuehl, M., Perrig, W.J., & Meier, B. (2010). The concurrent validity of the N-back task as a working memory measure. *Memory*, *18*, 394-412. doi: 10.1080/09658211003702171

Johns, B.T., & Jones, M.N. (2015). Generating Structure From Experience: A Retrieval-Based Model of Language Processing. *Canadian Journal of Experimental Psychology, 15,* 1196-1961. doi: 10.1037/cep0000053.

Just, M., & Carpenter, P. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*, 122–149. doi: 10.1037//0033-295X.99.1.122

Kamide, Y., Scheepers, C., & Altmann, G.T.M. (2003). Integration of Syntactic and Semantic Information in Predictive Processing: Cross-Linguistic Evidence from German and English. *Journal of Psycholinguistic Research, 32*, 37-55. doi: 10.1023/A:1021933015362

Kane, M.J. (2005). Full frontal fluidity? Looking in on the neuroimaging of reasoning and intelligence. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 141-163). Thousand Oaks, CA: Sage.

Kane, M.J., & Engle, R.W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin & Review*, *9*, 637-671. doi: 10.3758/BF03196323

Karimi, H., & Ferreira, F. (2016). Good-enough linguistic representations and online cognitive equilibrium in language processing. *The Quarterly Journal of Experimental Psychology*, *69*, 1013-1040. doi: 10.1080/17470218.2015.1053951

Kim, J.H., & Christianson, K. (2013). Sentence Complexity and Working Memory Effects in Ambiguity Resolution. *Journal of Psycholinguistic Research*, *42*, 393-411. doi: 10.1007/s10936-012-9224-4

Kiparsky, P. (2013). Towards a null theory of the passive. *Lingua, 125*, 7-33. doi: 10.1016/j.lingua.2012.09.003

Konieczny, L. (2000). Locality and parsing complexity. *Journal of Psycholinguistic Research*, *29*, 627-645. doi: 10.1023/A:1026528912821

Levy, R. (2007). Expectation-based syntactic comprehension. *Cognition*, *106*, 1-53. Doi: 10.1016/j.cognition.2007.05.006

Love, T., & Swinney, D. (1996). Coreference processing and levels of analysis in object-relative constructions: Demonstration of antecedent reactivation with the cross-modal

priming paradigm. *Journal of Psycholinguistic Research*, *25*, 5-24. doi: 10.1007/BF01708418

Mack, J.E., Meltzer-Asscher, A., Barbieri, E., & Thompson, C.K. (2013). Neural correlates of processing passive sentences. *Brain Sciences*, *3*, 1198-1214. doi: 10.3390/brainsci3031198

MacDonald, M.C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology, 4*, 226-242. doi: 10.3389/fpsyg.2013.00226

Maratsos, M.P., Fox, E.C., Becher, J., & Chalkley, M.A. (1985). Semantic restrictions on children's passives. *Cognition*, *19*, 167–191. doi: 10.1016/0010-0277(85)90017-4

Messenger, K., Branigan, H.P., McLean, J.F., & Sorace, A. (2012). Are children's early passives semantically constrained? Evidence from syntactic priming. *Journal of Memory and Language*, *66*, 568-587. doi: 10.1016/j.jml.2012.03.008

Osterhout, L., & Swinney, D. (1993). On the temporal course of gap-filling during comprehension of verbal passives. *Journal of Psycholinguistic Research*, *22*, 273-286. doi: 10.1007/BF01067834

Phillips, C., & Parker, D. (2014). The psycholinguistics of ellipsis. *Lingua*, *151*, 78-95. doi: 10.1016/j.lingua.2013.10.003

Reali, F., & Christiansen, M. H. (2007). Processing of relative clauses is made easier by frequency of occurrence. *Journal of Memory and Language*, *57*, 1–23. doi: 10.1016/j.jml.2006.08.014

Roberts, R., & Gibson, E. (2002). Individual differences in working memory. *Journal of Psycholinguistic Research, 31*, 573–98. doi: 10.1016/j.neuroscience.2005.07.002

Santi, A., & Grodzinsky, Y. (2010). fMRI adaptation dissociates syntactic complexity dimensions. *Neuroimage*, *51*, 1285-93. doi: 10.1016/j.neuroimage.2010.03.034

Shanks, D.R., Vadillo, M.A., Riedel, B., Clymo, A., Govind, S., Hickin, N., Tamman, A.J., & Puhlmann, L.M. (2015). Romance, risk, and replication: Can consumer choices and risk-taking be primed by mating motives? *Journal of Experimental Psychology: General*, *144*, 142-58. doi: 10.1037/xge0000116

Snyder, W. & Hyams, N. (2015). Minimality effects in children's passives. In E. Di Domenico, C. Hamann & S. Matteini (eds.) *Structures, Strategies and Beyond* (pp. 343-368). Amsterdam: John Benjamins.

Sprouse, J. (2007). Continuous Acceptability, Categorical Grammaticality, and Experimental Syntax. *Biolinguistics*, *1*, 118-129. doi: 10.1.1.692.1227

Sprouse, J., Wagers, M., & Phillips, C. (2012). A test of the relation between working-memory capacity and syntactic island effects. *Language*, *88*, 401-407. doi: 10.1353/lan.2012.0029

Staub, A. (2010). Eye movements and processing difficulty in object relative clauses. *Cognition*, *116*, 71-86. doi: 10.1016/j.cognition.2010.04.002

Street, J.A., & Dąbrowska, E. (2010). Lexically-specific knowledge and individual differences in adult native speakers' processing of the English passive. *Applied Psycholinguistics*, 1-22. doi: 10.1017/S0142716412000367

Swets, B., Desmet, T., Hambrick, D.Z., & Ferreira, F. (2007). The role of working memory in syntactic ambiguity resolution: A psychometric approach. *Journal of Experimental Psychology*, *136*, 64–81. doi: 10.1037/0096-3445.136.1.64

Swets, B., Desmet, T., Clifton, C., & Ferreira, F. (2008). Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory and Cognition*, *36*, 201-216. doi: 10.3758/MC.36.1.201

Thothathiri, M., Kim, A., Trueswell, J.C., & Thompson-Schill, S.L. (2012). Parametric effects of syntactic-semantic conflict in Broca's area during sentence processing. *Brain and Language*, *120*, 259-264. doi: 10.1016/j.bandl.2011.12.004

Townsend, D.J., & Bever, T.G. (2001). *Sentence Comprehension: The Integration of Habits and Rules*. Cambridge, MA: MIT Press.

Traxler, M.J., Corina, D.P., Morford, J.P., Hafer, S., & Hoversten, L.J. (2014). Deaf readers' response to syntactic complexity: Evidence from self-paced reading. *Memory & Cognition*, *42*, 97–111. doi: 10.3758/s13421-013-0346-1

Traxler, M.J., Morris, R.K., & Seely, R.E. (2002). Processing subject and object relative clauses: Evidence from eye movements. *Journal of Memory and Language*, *47*, 69-90. doi: 10.1006/jmla.2001.2836

Tutunjian, D., & Boland, J.E. (2008). Do We Need a Distinction between Arguments and Adjuncts? Evidence from Psycholinguistic Studies of Comprehension. *Language and Linguistics Compass, 2,* 631–646. doi: 10.1111/j.1749-818x.2008.00071.x

Van Dyke, J.A., & McElree, B. (2011). Cue-dependent interference in comprehension. *Journal of Memory and Language*, *65*, 247-263. doi: 10.1016/j.jml.2011.05.002

Volpato, F., Verin, L., & Cardinaletti, A. (2015). The comprehension and production of verbal passives by Italian preschool-age children. *Applied Psycholinguistics*, 1-31. doi: 10.1017/S0142716415000302